Mitglied der Helmholtz-Gemeinschaft

JÜLICH
FORSCHUNGSZENTRUM

Big Data at JSC

Platforms and Activities

January 15, 2015 | Björn Hagemeier

Integration

Develop an easily accessible big data platform

- Make HDFS available as a UNICORE storage
- Send YARN jobs through UNICORE to work on above storage

hadoop

UNICORE

Requirements

For our researchers, it would be ok to run in the cloud, but they need sufficient number of sufficiently large nodes to do real-world applications.
NASA did a test with weather data (MERRA/NEXRAD)

- 36 (30+6) nodes à 16 cores (2x8 Sandy Bridge) (=576 cores)
- 32GB RAM per node
- 36TB disk per node
- JSC cloud has 244 cores and 15TB block storage total

The complte datased was roughly 3.4TB.
Who in FedCloud could offer such ressources simultaneously?

Problems of BD Platforms

... for typical HPC environments

Assumptions that are “difficult” in an HPC center

- (almost) exclusive use of entire cluster
- direct, physical access to nodes’ local disks
- high RAM per node
- storage distributed among “compute” nodes
- BD software platform decides where jobs are run

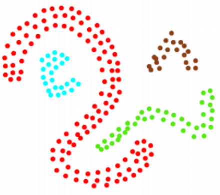
In summary

> *A big data platform is best made up of a dedicated pair of (appropriate) hardware and BD software platform. A “dedicated” cluster in the cloud is a viable solution.*

HPC versions of BigData algorithms

Parallel version of DBSCAN[1]

- DBSCAN is a clustering algorithm
- created higly parallelizable version (HPDBSCAN)
- tested @JSC with 3.7M points
- compared to alternative implementation
  - very good in terms of memory consumption, overall runtime, and speed-up
- Applications
  - Noise reduction
  - Twitter tweet density centers
  - Outlier detection



[1] http://en.wikipedia.org/wiki/DBSCAN

Spark and SciDB

This work is conducted in order to compare the NASA use case mentioned before on several big data platforms.

- Run on a real HPC cluster, but difficult to deploy
- Can use HDFS or local file system, but ...
- ... local file system on an HPC cluster is not really local